Preventing Fatal Car Crashes in Indiana

Charley Conroy, Chamreoun Chhay, Aziz Al Mezraani

Course: Time Series Forecasting

Spring 2025

Introduction:

Safety on the road is a life-or-death matter. A driver's license is a contract between an individual and their government that certifies that each driver in a given area is competent, responsible, and follows the law. Despite this, thousands of people a year are killed in car accidents, and although this number decreases as cars get safer, there's a lot that the government can do to minimize the risk of accidents.

The state of Indiana has a relatively low fatality rate when it comes to all 50 states, but it still underperforms when compared to its neighboring states. Compared with Michigan, Illinois, and Ohio, Indiana has a low number of fatalities but a high fatality rate per capita. Indiana has also been improving at a slower rate than some of these states. We wanted to take a closer look at the data to answer these questions:

1. What is the best model/method for future forecasts?
2. Can we accurately predict the crash rate for 2020 to 2022?
3. How will our model perform predicting up to 2026?
4. Do neighboring states experience similar fatality rates?
5. What policy changes can Indiana implement to lower crashes and fatalities?

To answer the questions above, we looked at the National Highway Traffic Safety Administration's car crash data for Illinois, Indiana, Michigan, and Ohio for the years 1999 to 2019 after determining that the data was not a random walk. For our predictions, we focused on just the Indiana data and revisited the other states for our later questions. We originally did not try to predict the COVID-19 pandemic years of 2020-2022 as that data was a random walk. We then tried to forecast those years, along with a forecast up to the end of 2025. After that, we took a look at the other states to see how Indiana performed and then researched state policies to try to offer the state of Indiana some policy recommendations to reduce car crash fatalities. This report contains our findings and some insights into how well our data has fared.

Selected States:

The NHTSA dataset had all 50 states in the United States, as well as data for the District of Columbia and the entire United States. We decided to use data for Indiana primarily because it's where all members currently reside, and we'd like to improve the safety of our state. Additionally, we decided to bring in data from 3 other neighboring states in the Great Lakes region of the Midwest: Illinois, Michigan, and Ohio. While we could have gotten data for states from other regions of the United States, we decided on these states because all 4 states have similar population sizes, weather patterns, laws, and driving conditions.

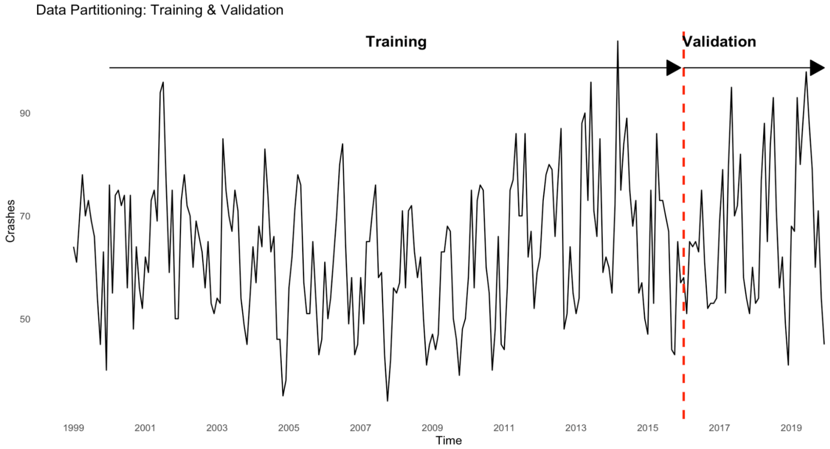
It is worth noting that Indiana has the lowest population by a decent margin with only 6,892,120 people compared to Michigan's 10,041,200 people, Ohio's 11,812,200 people, and

Illinois's 12,516,900 people [3]. We decided that these would still be worth considering, given that all 4 states are in the top 20 out of 50 states and the other factors that make these states valuable comparisons to Indiana.

Data Collection and Exploration:

Our goal was to build predictive models that use historical data on car crashes and weather conditions as supplemental data. For our analysis, we set the frequency to be monthly. Below are the two sets of data we used:

Set 1: The first dataset consisted of the monthly number of fatal car crashes in the state of Indiana from January 1999 to December 2022. This data was obtained from the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS). Early in our exploratory data analysis, we found that our time series data is a random walk. This means that a naive model can perform as well as a complex model. From there, we found that the random walk was caused by the impact of COVID-19 on the behavior of drivers during that time. Removing the COVID-19 period (2020 to 2022) from the time series makes it possible to apply more sophisticated models to the prediction. We then had the split of 85:15 for the training and validation of our models. From this split, we will test various models to see which one performs the best on the validation set.

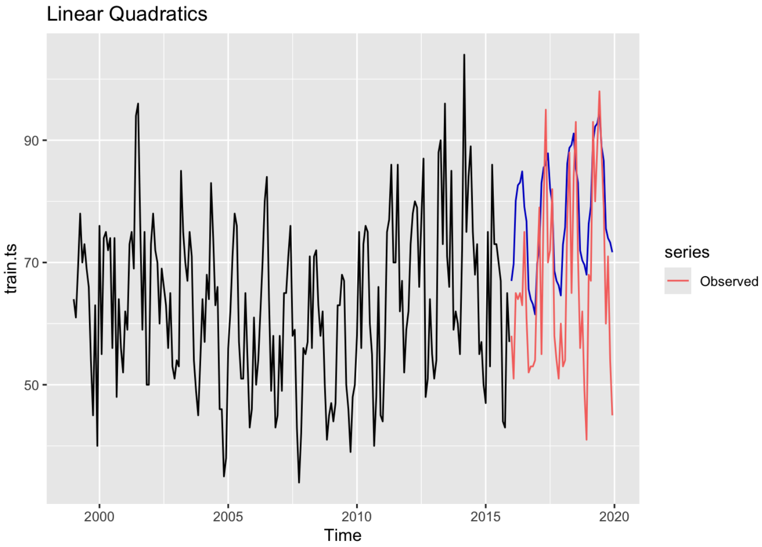


Set 2: The second data set contained weather data, specifically the average monthly temperature and average monthly precipitation for the same period, January 1999 to December 2022. This data is collected from Times Record News. The precipitation and temperature variables were included as potential explanatory variables in our predictive model to determine if they improve forecasting accuracy.

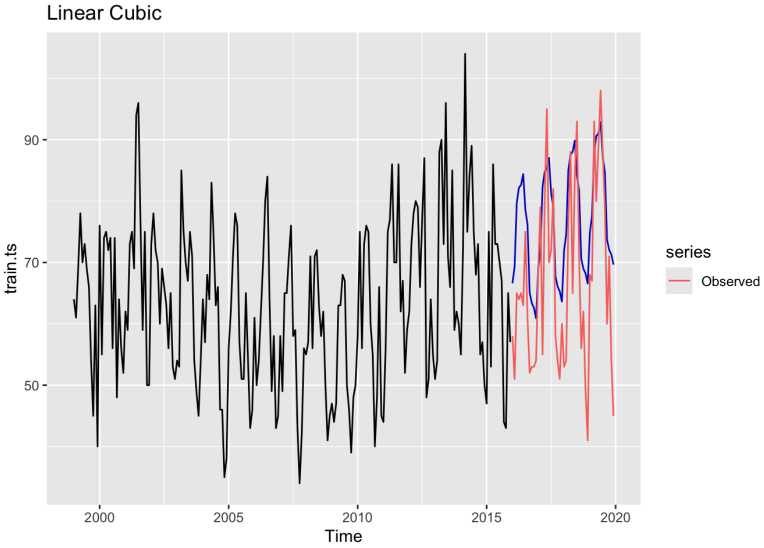
For the exploratory data analysis, we calculated the average number of car crashes for each month across the entire dataset. The number of crashes generally falls in the middle range, gradually increasing throughout the year, peaking around June, and then declining before reaching its lowest point in December (Figure 3). This seasonality gives us an idea of what could happen around that time of year.

Models Comparison

After looking at the training data, we realized a u-shaped curve. This pushed us to start our modelling process using a quadratic linear regression.

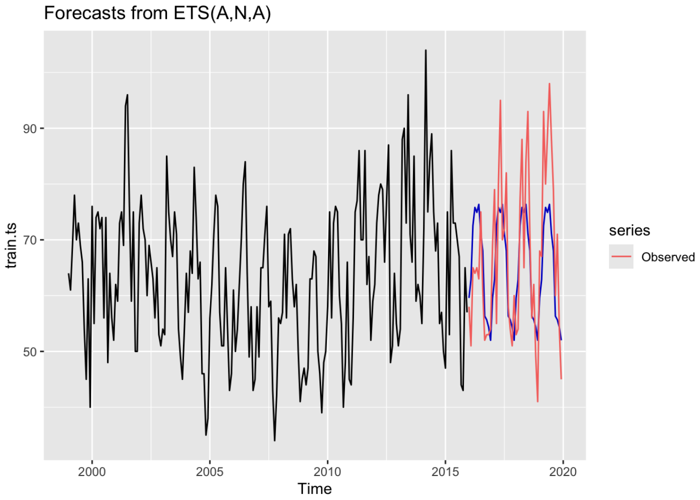


We can see that it was good at predicting the peaks and trend, but it was over-predicting the level and not catching any of the troughs. After looking a bit deeper, we realized that a cubic linear regression might be a better fit.

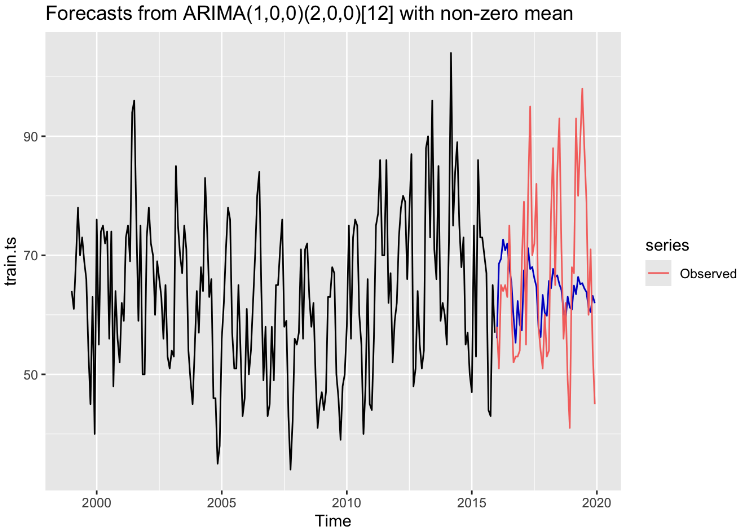


However, after running it, we realized that it was extremely similar to the quadratic regression with the same issues regarding the overprediction of the level and troughs while doing well with the peaks and trend.

Next, we decided to go into more complex models to see if they would capture the trend and seasonality better while being less affected by noise. The first complex model we ran was Holt-Winters.

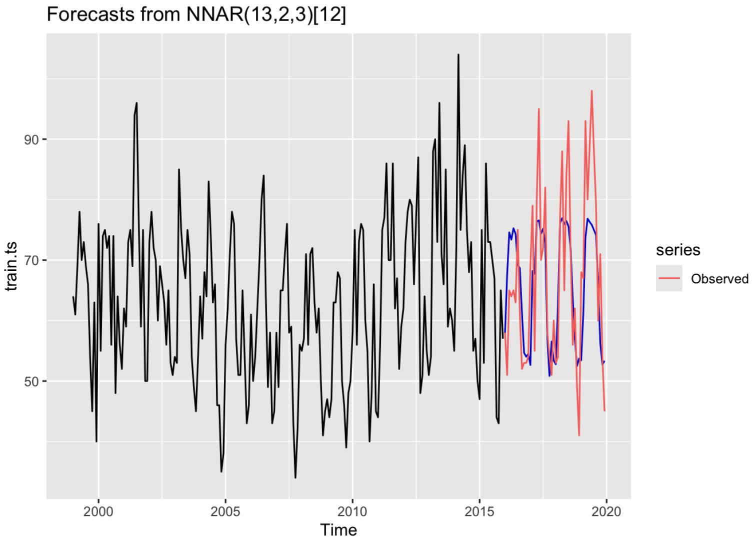


Using automatic parameters (additive errors and additive seasonality), we can see that while it does not catch the peaks and troughs very well, it captures the overall seasonality and trend fairly well. The second complex model we explored was an auto ARIMA model. The parameters selected ended up being 1 AR term and 2 seasonal AR terms with a seasonal period of 12 due to it being monthly data.



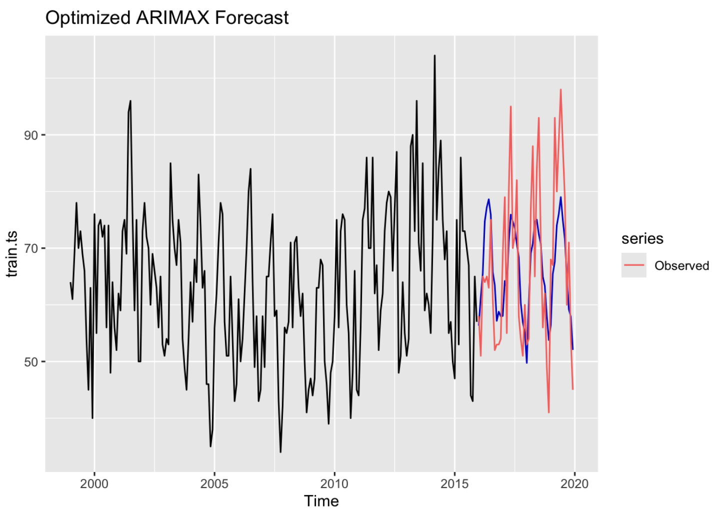
We can see that it caught the trend well, but it was not able to capture the seasonality, peaks, and troughs

We did 2 more complex models for which we actually used a grid search to find the optimal parameters based on the residuals of the forecast on the validation accuracy. The first one of those was the neural net. After running the grid search, we found that the optimal parameters are 13 lagged observations, 2 hidden layers, and 3 neurons for each layer, with the seasonal period being 12.



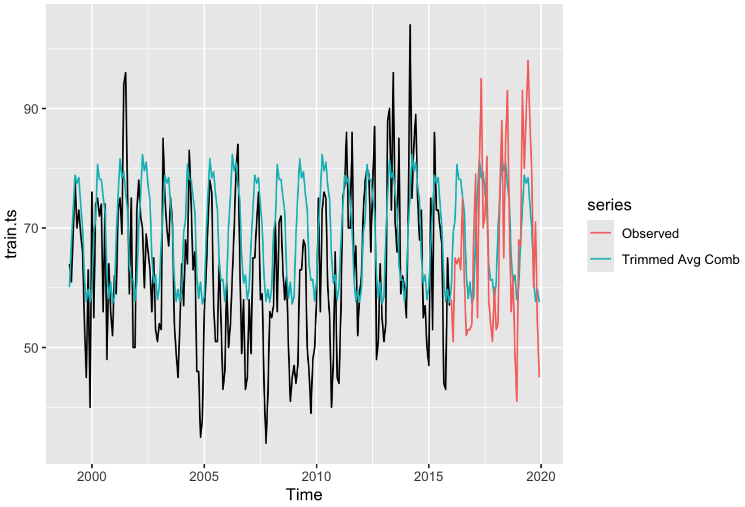
The prediction seems to fit well except for the peaks and troughs. However, it captures the level, trend, and seasonality well.

The last complex model we tested was an Arimax model with exogenous variables. We used temperature and rain data, hoping it would help find a relationship between the crashes and when the roads can be more dangerous due to weather conditions. (Include some info that it was the closest we could get.) After running the grid search, the recommended parameters were 1 AR term, first order differencing, 2 moving average terms, 2 seasonal MA terms, with a seasonal period of 12.



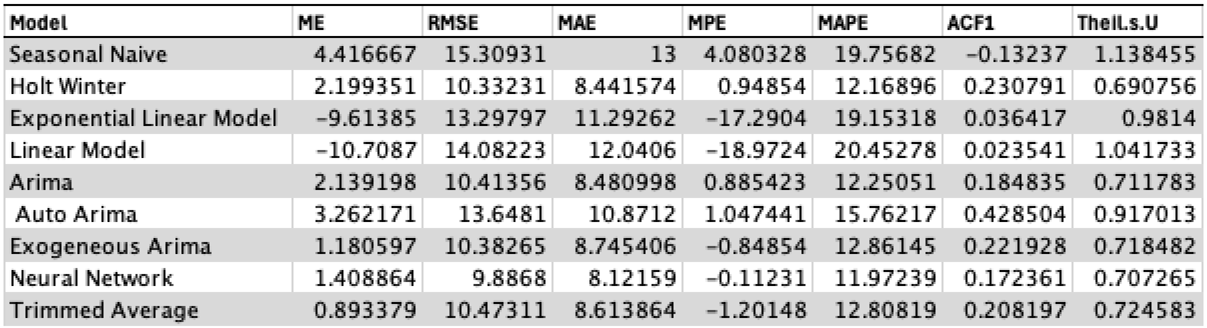
We can see that it predicted fairly well. It is similar to the neural network, but it overpredicted the level of the early years and was closer to the last peak, but it didn't capture the drops as accurately as the neural network.

We did one last forecast, which was an ensemble forecast using the trimmed average technique. This averages the performance of all models except those that had the lowest and highest prediction for each month.



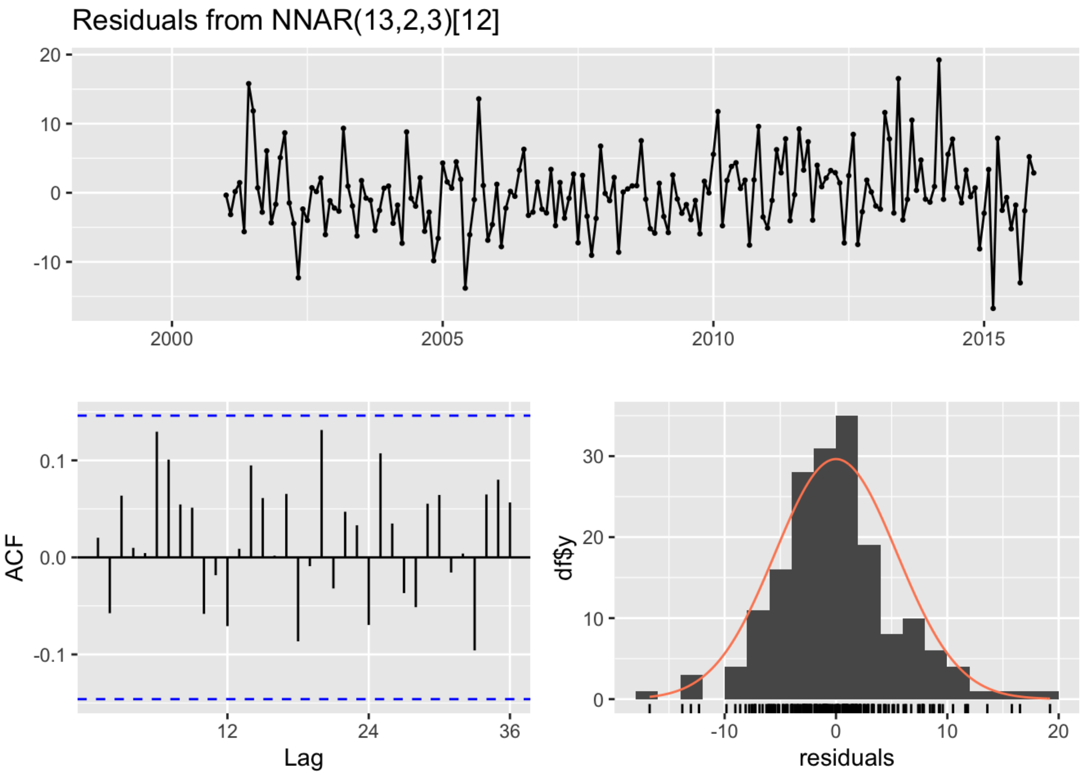
We can see that it did fine, but it doesn't seem to capture peaks and troughs as well as some of the other models.

To help us decide which model we would use for forecasting, we decided to compare their error measures (mainly RMSE and MAPE as those are the standard).



Model	ME	RMSE	MAE	MPE	MAPE	ACF1	TheilLs.U
Seasonal Naive	4.416667	15.30931	13	4.080328	19.75682	-0.13237	1.138455
Holt Winter	2.199351	10.33231	8.441574	0.94854	12.16896	0.230791	0.690756
Exponential Linear Model	-9.61385	13.29797	11.29262	-17.2904	19.15318	0.036417	0.9814
Linear Model	-10.7087	14.08223	12.0406	-18.9724	20.45278	0.023541	1.041733
Arima	2.139198	10.41356	8.480998	0.885423	12.25051	0.184835	0.711783
Auto Arima	3.262171	13.6481	10.8712	1.047441	15.76217	0.428504	0.917013
Exogeneous Arima	1.180597	10.38265	8.745406	-0.84854	12.86145	0.221928	0.718482
Neural Network	1.408864	9.8868	8.12159	-0.11231	11.97239	0.172361	0.707265
Trimmed Average	0.893379	10.47311	8.613864	-1.20148	12.80819	0.208197	0.724583

Looking at the values, the Neural Network seems to have performed the best on the validation set, thus, we decided to go ahead with it. Before we implemented it, we decided to explore if the residuals are capturing anything that the model is not.


Residuals from NNAR(13,2,3)[12]

Looking at these residual graphs, we can see that the errors fluctuate around 0 with some spikes. For the ACF plot, none of the lags pass the significance bans, which means there is no clear significant autocorrelation. As for the histogram, we can see that the spread is close to normal with a little skewness, However, nothing is alarming enough to justify running an AR(1) model on the residuals to help capture missing information from the model. Thus, the neural network alone is our final model.

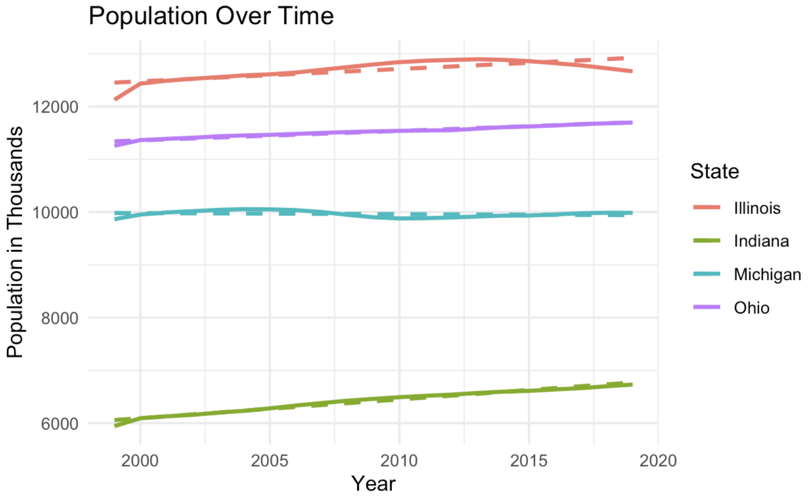
Model Implementation:

After deciding on our best performing model, we would employ a neural network due to it having the lowest RMSE and MAPE. We would first train the model using historical crash data up to 2022, testing its accuracy by comparing predicted values with actual recorded values, allowing us to assess how well our model captures patterns and trends in accident rates. Given that the COVID-19 pandemic had an impact on road usage, with fewer cars on the road, we would then analyze the deviation in our model's predictions compared to actual crash rates during this period. This comparison helps us determine whether the pandemic caused a temporary anomaly in crash trends and whether our model effectively recognizes this shift.

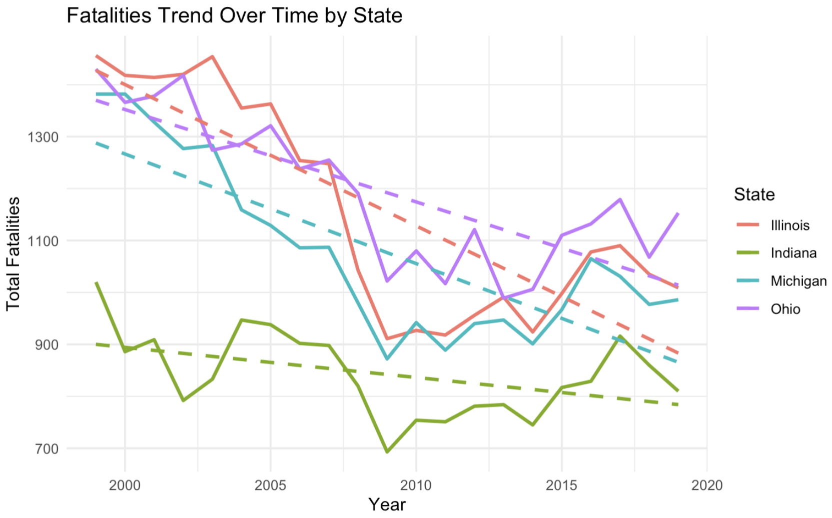
After identifying any discrepancies caused by the pandemic, we would retrain the model by incorporating the COVID-19 period back into the dataset. By including data from 2020 to 2022, we ensure the model learns from real-world disruptions, making future forecasts more reliable. This updated model would then be used to predict crash rates up to 2026, capturing both long-term trends and potential improvements in road safety (Figure 4). By leveraging neural networks, we can generate more nuanced forecasts, identifying whether crash rates are improving or if they follow pre-pandemic patterns.

Great Lakes States Comparison:

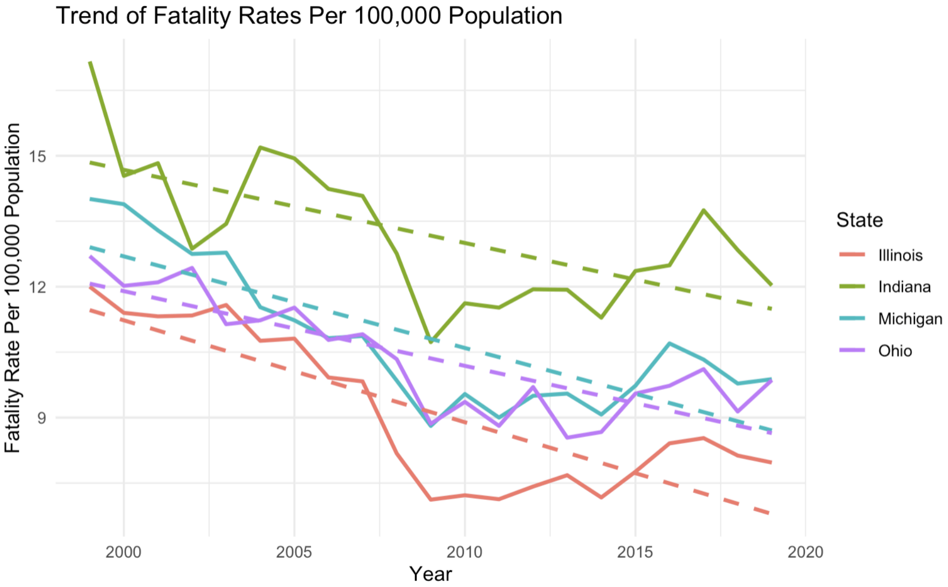
The first thing we had to consider when comparing the Great Lakes states was that the state of Indiana has had consistently lower populations than the other 3 states, as shown in the below graph.



One thing to note about this is that although the population is so much lower than the other states, the total number of car crash fatalities in Indiana has not come down nearly as much as its neighbors, as seen in the graph below. Michigan, Illinois, and Ohio have significantly closed the gap between them and Indiana over time despite having nearly double the population.



This is concerning enough on its own, but looking at fatalities paints a far more grim picture, as seen below. Indiana has the highest fatality rate adjusted for population. The Indiana government has been failing compared to its neighboring states when it comes to keeping its citizens safe on the road.



This all implies that the other states have been making changes that Indiana has not, and that Indiana should be considering what policies they can implement from these other states.

Recommendations:

We took a look at the policies for all 4 states from the lens of reducing fatal crashes from 2 high-risk groups: new drivers and drunk drivers.

Indiana does have some policies that are more strict than the other states, specifically when it comes to drunk driving. For instance, Indiana is the only state that has Operating While Intoxicated (OWI) as the charge instead of just Driving Under the Influence (DUI). What this means is that operating a vehicle at all or even turning on a vehicle while intoxicated is illegal, which is a much stricter law than needing the offending vehicle to be in motion and applies to all vehicles, including bicycles.

Additionally, Indiana is the only state without medical or recreational marijuana legalized, making driving while under the influence of marijuana carry additional charges for possession of an illicit substance. Finally, their OWI penalty can carry license suspension for up to 2 years for first-time offenders, which is the highest of the 4 states [].

However, Indiana falls behind in a few key areas for intoxicated driving, especially for the potentially lax nature of punishments. First-time offenders can get a 2-year license suspension, but they can also get a 30-day suspension, which is the lowest of the 4 states. This makes Indiana possess by far the most flexible punishment, as the other 3 states do not have ranges like this.

Additionally, while Ohio does not have Mandatory Alcohol Education as a punishment like Michigan and Illinois, Indiana has Mandatory Alcohol Education or Assessment and Treatment as optional punishments while the others do not. This makes first-time OWI offenders in Indiana carry the most lax penalty for intoxicated driving, which could encourage irresponsible drivers to offend or recidivate. All this makes Indiana a dangerous place to drive, especially during the summer months when drunk driving can double [3].

For new drivers, Indiana is close to the other states but still lags when it comes to supervised driving and behind-the-wheel training. Illinois has similar policies, which are lower than Michigan and Ohio, but has an additional 50 hours of supervised driving. Requiring supervised driving and additional behind-the-wheel training could better prepare young drivers to be safe on the road by letting more experienced drivers offer advice for more niche situations. Additionally, splitting the 30 hours of class time into 2 segments like Michigan could be more effective at getting new drivers to retain the information.

Ultimately, with a few policy changes to assist Indiana's newest drivers and punish reckless drivers, Indiana could bring down its car crash fatalities to be among the lowest in the country. Additionally with data closer to the present we could better approximate how effective these changes will be at preventing car crash fatalities in Indiana.

References

[1] Administration, National Highway Traffic Safety. “Fars Encyclopedia.” *FARS Encyclopedia: Crashes - Time*, www-fars.nhtsa.dot.gov/Crashes/CrashesTime.aspx. Accessed 28 Feb. 2025.

[2] Amir. “State-by-State Guide to Driver’s Education Requirements | Drive Rite NY.” *Driveriteenv.Com*, driveriteny.com/Blog/State-by-State-Guide-to-Drivers-Education-Requirements. Accessed 28 Feb. 2025.

[3] Keller “Drunk Driving Accidents Increase in Summer Months.” *Keller & Keller*, www.2keller.com/library/how-drunk-driving-accidents-increase-in-the-summer-months.cfm. Accessed 28 Feb. 2025.

[4] *US States - Ranked by Population 2024*, worldpopulationreview.com/states. Accessed 28 Feb. 2025.

[5] Vandervort-Clark, Amy. “State-by-State DUI Penalties - Findlaw.” *Findlaw.Com*, www.findlaw.com/dui/laws-resources/state-by-state-dui-penalties.html. Accessed 28 Feb. 2025.

[6] Times Record News. “Weather Data - Indiana.” *Data Times Record News*, data.timesrecordnews.com/weather-data/indiana/18/1987-12-01/table/. Accessed 28 Feb. 2025.

Appendix

Figure 1

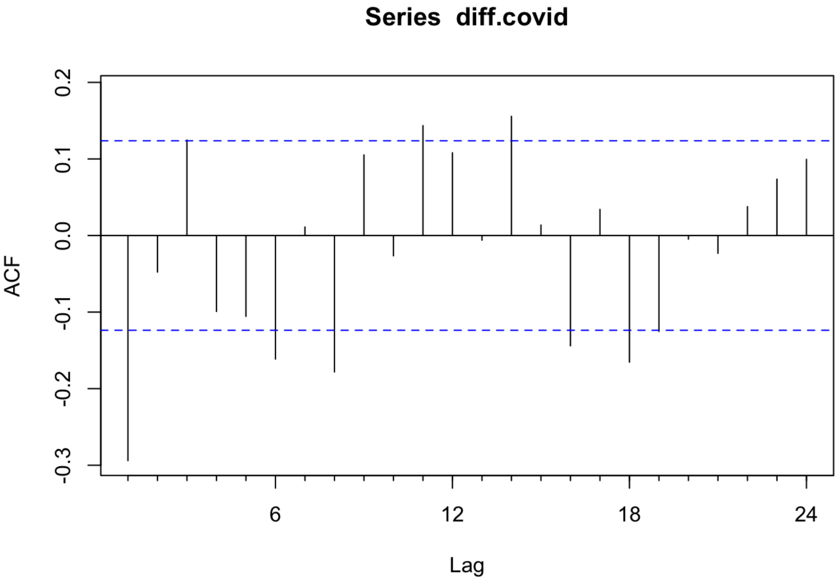


Figure 2

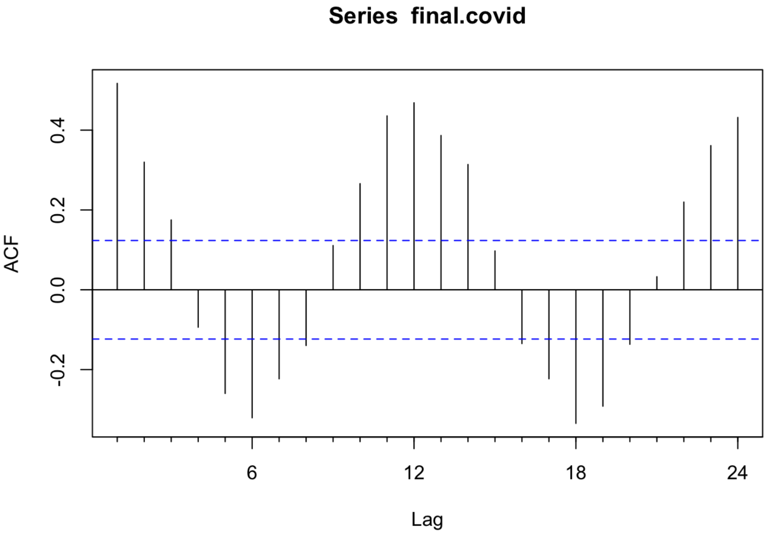


Figure 3

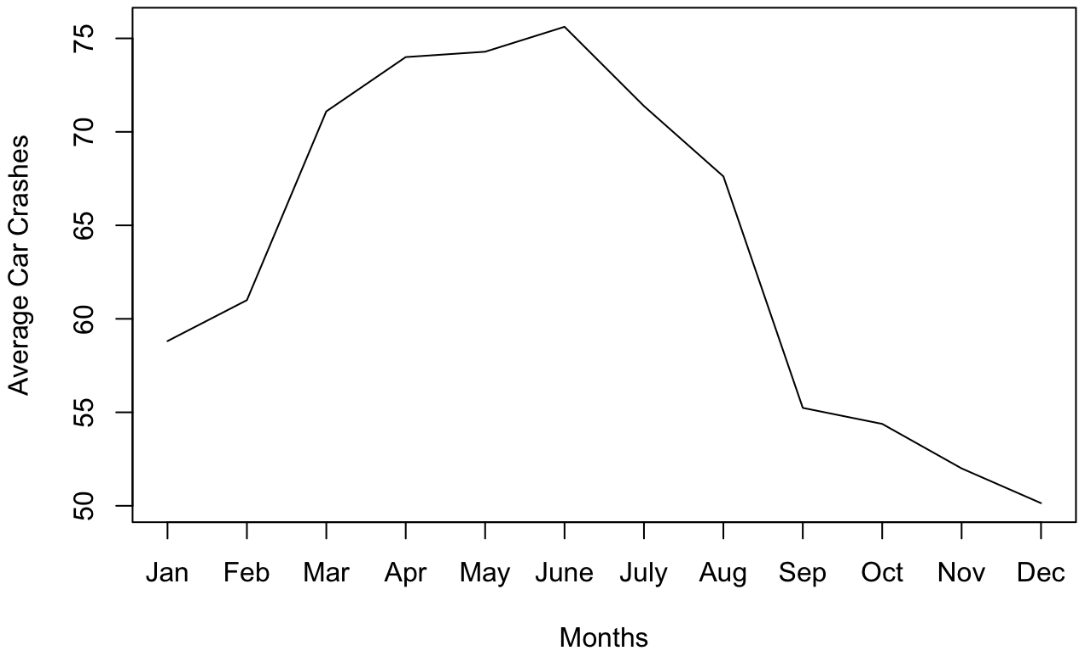
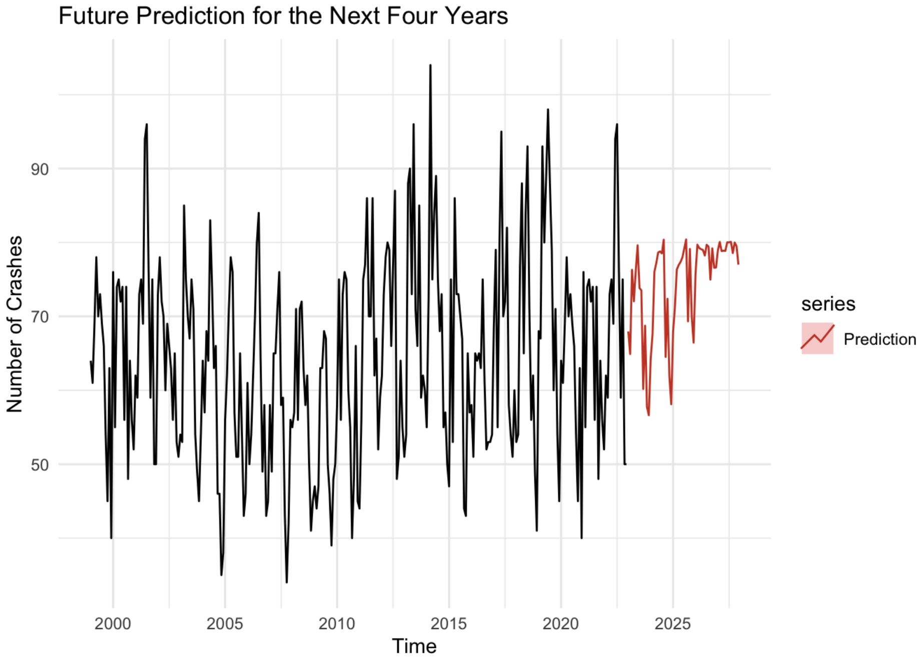


Figure 4


Future Prediction for the Next Four Years
Number of Crashes
Time
series
Prediction